Inter-method Performance Study of Tumor Volumetry Assessment on Computed Tomography Test-retest Data

Andrew J. Buckler, MS[1]
Jovanna Danagoulian, PhD,[1]
Kjell Johnson, PhD,[2]
Adele Peskin, PhD,[3]
Marios A. Gavrielides, PhD,[4]
Xiaonan Ma, MS,[1]
Maria Athelogou, PhD,[5]

[1]Elucid Bioimaging Inc., 225 Main Street, Wenham, MA 01984
[2]Arbor Analytics LLC, 4079 Ramsgate Court, Ann Arbor, MI 48105
[3]National Institute of Standards and Technology, 325 Broadway, Boulder, CO 80305
[4]U.S. Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993
[5]Definiens AG, Bernhard-Wicki-Straße 5, 80636 München, Germany

ABSTRACT

Thirty-one lung cancer test-retest cases were analyzed by twelve participants in a multi-site study of algorithm performance on the segmentation of clinical CT scans. We evaluated variability of scalar volume measurements, including individual participant performance across test-retest repetitions, as well as the performance across algorithms. We also compared segmentation boundaries relative to reference standard segmentations. We report the repeatability of measurements in test-retest cases for each participant, and reproducibility of measurements across participants. Repeatability coefficients (RC) ranged from .06 (best performing) to 1.5 (least performing), corresponding to within-subject coefficients of variation of 2.1% to 54% respectively. Reproducibility coefficient values (RDC) are somewhat greater than the lowest performing pooled participant's repeatability value. The best algorithm performance is seen when measured tumors meet the measurability criterion defined in the QIBA Profile; including tumors that did not resulted in approximately 1-1/2 times the variability. The value of editing segmentation results was equivocal; smaller tumors appear to be better without editing, but larger tumors benefit by editing. Linear mixed effects modelling was used to conclude that no more than two-thirds of the overall QIBA Profile variability claim of the system as a whole be allocated to analysis software if the overall system is to be compliant (or less if conditions such as the scanner settings are not held constant). An important outcome of this work is that the set of metrics used for this analysis form a basis for future determination of compliance with the QIBA Profile.

I. INTRODUCTION

Quantifying tumor volume change is being studied for use as an imaging biomarker with application to diagnosis, therapy planning and evaluating response to therapy. A biomarker is defined generically as an objectively measured indicator of a normal or pathological process or pharmacologic response to treatment [1, 2]. A quantitative imaging biomarker is defined as a measurand where each of the following is true: 1) the difference between two measurements is meaningful, and 2) there is a clear definition of zero such that the ratio of two measurements is meaningful [3, 4]. The use of tumor volume as a predictor of outcome has been of interest for some time [5-7]. A number of authors have reviewed the use of tumor volumetry using computed tomography (CT) and how it has evolved [8-12]. A number of published studies investigate the link between tumor volume at CT and cancer disease status [13-23].

On the technical level, biomarker assays need to be characterized in terms of bias and variability. Bias and variability in serial CT scans can be affected by a number of inter-related factors, including imaging parameters, tumor characteristics, and/or measurement procedures [15]. These effects have to be understood and quantified to establish confidence in the use of volumetric CT measurements for clinical study cohorts. A number of technical studies have been performed toward that goal [24-39].

The Quantitative Imaging Biomarker Alliance (QIBA) [40] has defined standard procedures for measuring lung tumor volume changes in a document called a Profile, which defines standard working procedures for accurate and reproducible measurement of imaging biomarkers. The Profile is defined in part by available literature, and in part by "groundwork" studies to investigate sources of error in volume estimation, for example, by different acquisition protocols, scanner models, etc. An important branch of study that QIBA has commissioned is an investigation into volumetric measurement algorithm performance, under the name "3A". The aim of the first QIBA 3A study was to estimate intra- and inter-algorithm bias and variability on phantom data sets. The study was organized as a public challenge where participant algorithms were applied to FDA acquired phantom CT scans of synthetic lung tumors in anthropomorphic phantoms [41]. Such a study design is effective for a focus on bias, since ground truth is known, but is likely to underestimate variability, since typically clinical data sets are more challenging due to a variety of biological and technical reasons. As a result, QIBA has also undertaken studies on clinical data, notably a study of human reader performance on test-retest data, under the name "1B". The 1B study was undertaken to determine the variability of lesion size measurements in CT datasets of patients imaged under a "no change" ("coffee break") condition and to determine the impact of two reading paradigms (independent readings of both time points vs. locked sequential readings) on measurement variability (publication in progress). This second 3A project is best understood as a complement to both the first 3A study by examining inter-algorithm variability for the measurement of lung tumors in clinical as opposed to phantom data sets, and the 1B study by using the same clinical data.

A collaborative approach was taken to design the challenge study, resulting in the *Study 3A: Inter-method Study with Test-retest Clinical Data: Study Design* document found at [42]. The challenge used CT scans of 31 lung cancer patients, scanned twice within 15 minutes and reconstructed as thin transverse slices. Twelve participants from a diverse set of industry and academic groups downloaded the images and supporting information and uploaded results, using a system called QI-Bench [43]. QI-Bench provides open-source informatics tools to characterize the performance of quantitative medical imaging, including support for accessing image archives, representation of meta-data, and computing statistical metrics consistent with QIBA's initiative in metrology.

Whereas ground truth for volume was not available to directly assess bias, this study was based on test-retest data, consisting of clinical scans of patients repeated close enough in time that no biological change could have taken place. As such, the repeatability of volume measurements in test-retest sets may be determined for each algorithm (intra-algorithm repeatability) as well as across algorithms (inter-algorithm reproducibility). Using actual segmentation results (not just the computed volumes) that were submitted by a subset of the participating groups, it was also possible to compare segmented boundary contours, providing further metrics to characterize performance and providing insight into the differences in algorithm performance. The metrics computed based on measured volumes taken together with the metrics available through analyzing the contours may serve as means to determine relative performance levels for use in determining whether a given algorithm may be said to be "QIBA compliant," understood as having performance at or better than requirements specified under the QIBA Profile [44]. Additionally, the results of this study provide experimental data for updating the Profile requirements themselves as our understanding of attainable performance is extended.

II. MATERIALS AND METHODS

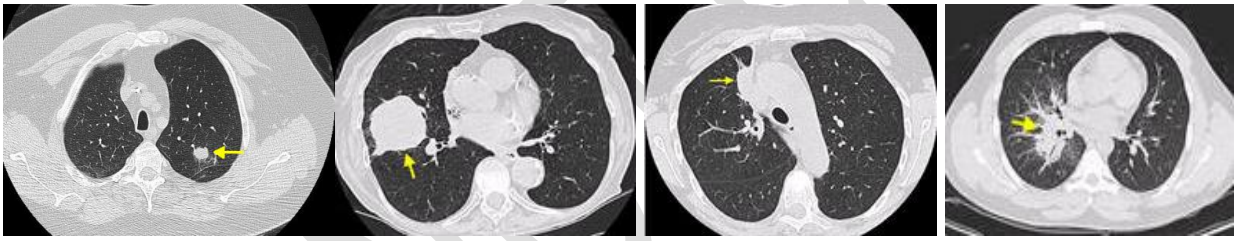
Data and Data Collection

Thirty-one non-small cell lung cancer test-retest cases were used in this analysis These cases were contributed to the RIDER database from Memorial Sloan Kettering [45], with a mean patient age of 62.1 years, range, 29–82 years; 16 were men (mean age, 61.8 years; range, 29–79 years) and 16 were women (mean age, 62.4 years; range, 45–82 years). Each patient was scanned twice within a short period of time (< 15 minutes) on the same scanner and the image data was reconstructed with thin sections (< 1.5 mm thick). CT scans were obtained with a 16–detector row (LightSpeed 16; GE Healthcare, Milwaukee, Wisconsin) or 64–detector row (VCT; GE Healthcare) scanner. Parameters for the 16–detector row scanner were as follows: tube voltage, 120 kVp; tube current, 299–441 mA; detector configuration, 16 detectors x 1.25-mm section gap; and pitch, 1.375:1. Parameters of the 64–detector row scanner were as follows: tube voltage, 120 kVp; tube current, 298–351 mA; detector configuration, 64 detectors x 0.63-mm section gap; and pitch, 0.984:1. The thoracic images were obtained without intravenous contrast material during a breath hold. Since the second scan was considered as a separate scan, its field of view was set

given the patient's second scout image. Adjustment was allowed owing to the patient's position in the scanner. Thin-section (1.25 mm) images were reconstructed with no overlap by using the lung convolution kernel and transferred to the research picture archiving and communication system server where Digital Imaging and Communications in Medicine (DICOM) images are stored.

One tumor per patient was selected by clinical staff at Columbia University for measurement (31 tumors total), and each tumor had two repetitions ("test" and "retest"). The approximate diameters ranged from 8 mm to 40 mm. The shapes of the selected tumors ranged from simple and isolated to complex and cavitated. To facilitate the comparison of results with the prior QIBA 1B study, the tumors were further subdivided on the basis of "measurability" criteria as described in the QIBA Profile. Specifically, the claims section of the QIBA profile states that the claims are only applicable "*when the given tumor is measurable (i.e. tumor margins are sufficiently conspicuous and geometrically simple enough to be recognized on all images….) .and the longest in-plane diameter of the tumor is 10 mm or greater*". Therefore, tumors described as meeting the QIBA Profile were those that were judged to have clearly identified tumor margins; all tumors used in this study exceeded the 10 mm diameter threshold.

Illustrative examples are given in Figure 1.



(a) (b) (c) (d)

Figure 1: Examples of tumors used. (a) and (b) are examples of tumors that were judged to have met the QIBA Profile, while (c) and (d) were examples of tumors that were judged to have not met the QIBA Profile.

RSNA staff handled participant agreements and communications so as to establish and maintain anonymity of participants with respect to the results. Participants downloaded the challenge data from QI-Bench, including the raw image data as well as location points defined for each tumor in the scans. The location points were defined to lie within the tumor margin, but participants were allowed to select different or multiple seed point(s) for their individual algorithms, provided they utilized the tumor identification scheme provided. Each participating group measured each tumor at each repetition. Some of the groups submitted data from the algorithm without modification (fully automatic), others submitted data that had been adjusted to varying degrees by a reader (semi-automated), and one group submitted both (without post-editing under one group ID and adjusted under a separate group ID).

Two statistical analyses were conducted, based on the type of data: 1) variability of scalar volume measurements, including individual participant performance across test-retest repetitions as well as the performance across algorithms, and 2) comparison of segmentation boundaries

relative to reference standard segmentations. The former allows us to compare the performances of these imaging algorithms by measuring agreement of the computed result when the algorithm is held constant as well as when measured by different algorithms, regardless of the similarity in the contours that give rise to the scalar volumes; the latter provides the means by which differing algorithms may be evaluated in terms of the specific segmentation task they are performing which gives rise to the computed scalar volumes.

Variability of Scalar Volume Measurements

Data transformation: The models used in our analysis assume a constant variance in volume measurements across the range of the responses. Since measurement variation was not constant across the range of volumes and increased with increasing volume measurements, volume measurements were transformed so that the constant variance assumption would hold. In order for residual values (the differences between measured volumes and mean volumes calculated from the set of algorithms) to be the same order of magnitude for all tumor sizes, a log-transformation was applied to each volume. As a result, residuals approximately followed a normal distribution. Although analyses were conducted on the log-scale, data is presented on the original scale, where possible.

Based on the transformed data, we undertook two analyses of volume measurement variability in this study, *repeatability* and *reproducibility* [3] using visual as well as numeric methods. Plotting test-retest replications (for repeatability) or pair-wise combinations of algorithms (reproducibility) appear as a straight line of unity in the presence of agreement. Numerically, we denote the measurement of the j^{th} algorithm for the i^{th} subject at the k^{th} replication as Y_{ijk}, where j=1,…,11, i=1, …,31, and k=1, 2. We used a simple general model $\boldsymbol{Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}}$, where Y_{ijk} and ε_{ijk} are the observed value and measurement error and where μ is the population mean. $\boldsymbol{\mu_{ij}}$ is conditional on the mean of infinite replications made on subject i by algorithm j. Based on these analyses, we compute multiple metrics because each provides complementary insight into performance.

Repeatability across test-retest repetitions within participants, or intra-algorithm variability, refers to the variability of measuring the volume of the same tumor from repeated imaging of subjects with intentionally short interval so that biological features could be reasonably assumed to have remained unchanged. The Bland and Altman method produces an Upper Agreement Limit (*UAL*) and the Lower Agreement Limit (*LAL*) which provides a range within which we expect 95% of the differences between replicate measures of a given algorithm [46, 47]. The Concordance Correlation Coefficient (CCC) was used as a measure of repeatability, computed as in [48]. CCC is a measure of agreement that is a product of the correlation coefficient, penalized by a bias term that reflects the degree to which the regression line differs from the line of agreement. The further the regression line is from the line of agreement, the higher the penalty, and the lower the CCC. The repeatability coefficient (RC) is a function of the standard deviation of the measurements:

$$RC = 1.96\sqrt{2\sigma_\varepsilon^2} = 2.77\sigma_\varepsilon.$$

We define a range of measurements (-RC to +RC), in which two normally-distributed measurements are expected to fall for 95% of replicated measurements [49]. The within-subject standard of deviation (*wSD*) is estimated as square root of the averaged sample variances across tumors, where the sample variance is computed from the replications for each tumor. This *wSD* assumes that the within-tumor variance is the same across all tumors. The within-subject coefficient of variance (*wCV*) is a relative measure of repeatability, which we calculate as *wSD*/mean and thus is proportional to the magnitude of the tumor's size.

Reproducibility across algorithms was analyzed similarly but instead of the two repetitions, pairwise comparisons were made between algorithms. In this case, the Limits of Agreement (LOA) by Bland and Altman provides a range within which we expect 95% of the differences in measurements between two algorithms to lie. The reproducibility coefficient (RDC), similar to RC, was calculated as the least significant difference between two measurements taken under different conditions, in this case, by two different algorithms. Linear Mixed Effects (LME) modeling was used to separate the factors that affect variability. Each of these terms was considered as a random effect in the model. Model assumptions were evaluated with Q-Q (quantile-quantile) and observed-versus-fitted plots. We are interested in measuring to what extent algorithm versus other variance contributes to overall error, to better define the QIBA claim.

Comparison of Segmentation Boundaries

Whereas the nature of clinical data makes actual ground truth unavailable, we can approximate a reference segmentation using those pixels with the highest agreement among participants. We first produced a reference segmentation using the Simultaneous Truth And Performance Level Estimation (STAPLE) method [50]. This filter performs a pixel-wise combination of an arbitrary number of input images. In our case we use the segmentations performed by participant algorithms. Each input segmentation is weighted based on its "performance" as estimated by an expectation-maximization algorithm, described in detail in [51]. We then compare each individual segmentation result to this reference data. We compute Sensitivity (*SE*) or true positive rate, based on a confusion matrix *C*, where C_{uv} is the number of voxels segmented with a particular algorithm *u*, compared with the reference data *v*. For any label *w*, we calculate true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*FN*) as:

$$TP = C_{ww} \quad TN = \sum_{\substack{u=1 \\ u\neq l}}^{N} \sum_{\substack{v=1 \\ v\neq w}}^{N} C_{uv} \quad FN = \sum_{\substack{u=1 \\ u\neq w}}^{N} C_{uw} \quad FP = \sum_{\substack{v=1 \\ v\neq w}}^{N} C_{vw}.$$

$$SE = TN/(TN + FP).$$

Typically *SE* is accompanied by Specificity, otherwise known as the true negative rate. However, this quantity has a strong dependence on the size of the field of view which is constant for all participants so we omit reporting this as it is not informative. *TP* and *FN* computations

are used in the calculation of two additional spatial overlap measures, the Jaccard index [52], and Sørensen–Dice coefficients [53, 54]:

$$Jaccard = \frac{TP}{TP+FP+FN} \quad SørensenDice = \frac{2\times TP}{2\times TP+FP+FN}$$

As compared to Sensitivity, the Jaccard index penalizes false positives, i.e., if the candidate segmentation is too large because it includes anatomy not contained in the reference. Sensitivity would not pick this up, but the Jaccard index does. Sørensen–Dice not only does this, but weights the overall measure stronger than the others on true positives; i.e., it penalizes candidate segmentations that haven't picked up anatomy that is contained in the reference. While at some point it may be evident which is the more important, for this work we compute and present all three types of numeric comparisons, collectively described as "overlap metrics."

III. RESULTS

Twelve groups participated in the challenge by submitting volume readings and five of those groups also submitted segmentation objects, four of which were compatible for analysis. The following groups participated in the challenge study (sorted in alphabetical order rather than in numeric order of the IDs):

- *Fraunhofer MEVIS*
- *GE Healthcare*
- *ICON Medical Imaging*
- *KEOSYS*
- *MEDIAN Technologies*
- *Medical University of South Carolina*
- *Mirada Medical*
- *Perceptive Informatics*
- *Siemens AG*
- *UCLA*
- *University of Michigan*
- *Vital Images*

See the appendix for detailed algorithm descriptions for each of the participating groups. (Note that three groups (Group01, Group09, and Group13) initially applied but did not submit results, and Group10 and Group16 were synonymous IDs.)

The following subsections present the results of the analyses described above:

1. Variability of scalar volume measurements:
   1.1. Descriptive statistics and transformation.
   1.2. Repeatability across test-retest repetitions within participants, using numeric calculation and Bland-Altman analysis to report *UAL*, *LAL*, *wSD*, *RC*, *wCV*, and *CCC*.
   1.3. Reproducibility across participating algorithms:
      1.3.1. Consider heteroscedasticity to determine a set to be pooled;
      1.3.2. Analysis including all tumors / all algorithms:

- Numeric calculation and Bland-Altman analysis to report *UAL*, *LAL*, and *RDC*.
- LME model to assess subject, algorithm, and residual variance.

1.3.3. Stratified analyses based on "measurability" criterion defined in the QIBA Profile and another that partitions algorithms according to degree of post-editing.

2. Comparison of tumor segmentation boundaries:
   2.1. Creation of reference segmentation for each patient and each test-retest repetition.
   2.2. Calculation of sensitivity, Jaccard index, and Sørensen–Dice coefficients.
   2.3. Merging and plotting of histograms by metric and participant.

1. Variability of Scalar Volume Measurements

Our goal in this section of the results was to determine the repeatability in volume calculations across acquisitions in test-retest data for each participant, and to determine the reproducibility of these results across all participants.

1.1 Descriptive Statistics and Transformation

Basic descriptive statistics on submitted measurements are given in Table 1, based on measurements of 31 lung tumors at each of two repetitions by 12 participants, and a total of 744 measurements.

Table 1: Basic Descriptive Statistics for measured tumor volume

Basic Descriptive Statistics (mm^3)	
Arithmetic Mean	2.41E+04
Geometric mean	8.32E+03
Median	9.11E+03
Range	2.77E+05
Minimum	3.00E+00
Maximum	2.77E+05

Figure 2 summarizes the range and distribution of these readings. The distribution is skewed due to very few large reading values, also seen in Table 1, where the mean is much higher than the median.

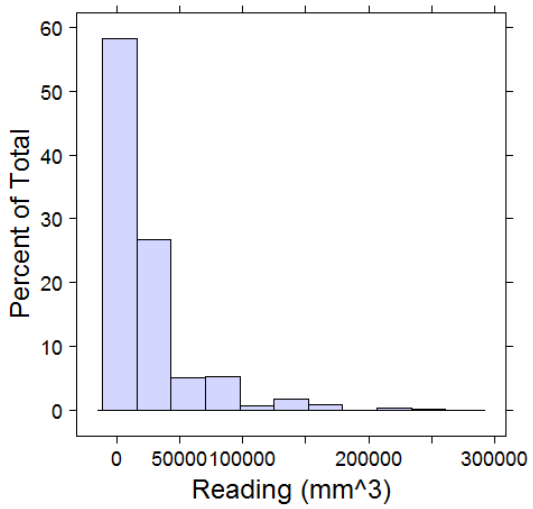


Figure 2: The distribution of measured values of tumor volume (Reading) across image acquisitions and participants.

1.2 Repeatability Across Test-Retest Repetitions Within Participants

Repeatability was assessed separately for each participating group since by definition it involves holding the algorithm constant. Individualized reports were prepared for each participant accordingly. Results for Group12 are given to illustrate the method, but each participant had different results. Variability of volume measurements is shown in Figure 3, first by plotting volumes from the first acquisition against those of the second. Like the histogram displayed in Figure 2, the distribution of volumes is skewed. Furthermore, this figure illustrates that the variation in the test and retest measurements increases with the mean. An assumption of a Bland-Altman analysis is that the variation is constant across the range of the response. This pattern indicates that a transformation of the data is necessary. The log-transformation is commonly applied to skewed data; this transformation helped the data to meet the assumptions of the analysis. The Bland-Altman plot is displayed on the right side of Figure 3, and the variation on the transformed scale is approximately constant across the range. The 95% limits of agreement on the log10 scale are -0.19 and 0.24. Consider, for example, a mean volume reading of 1000 mm^3 (or 3 on the log10 scale). The Bland-Altman limits of agreement would be between 646 and 1738 (or $10^{2.81}$ and $10^{3.24}$, respectively). Next, consider a mean volume reading of 100,000. The Bland-Altman limits of agreement here would be 64,565 and 173,780 (or $10^{4.81}$ and $10^{5.24}$, respectively).

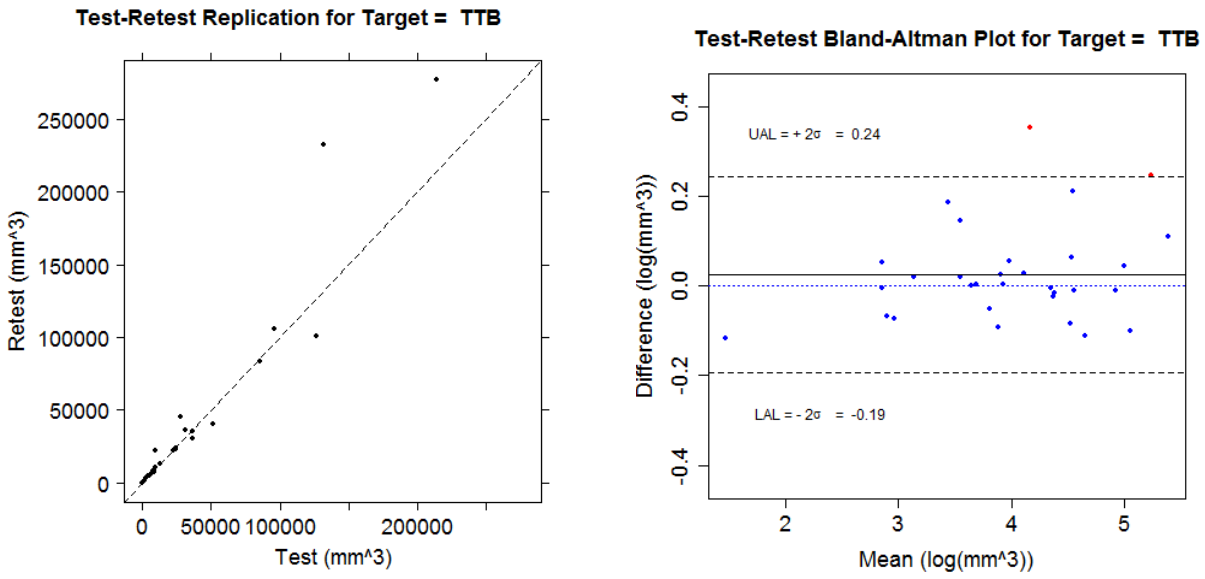


Figure 3: Results of Intra-algorithm analysis: Left panel: Scatter plot for test-retest conditions in the original units of measurement. Notice that the variation increases with the volume (labelled TTB). This artifact implies that the data need to be transformed prior to Bland-Altman analysis. Right panel: Bland-Altman chart on the log-transformed data. The Upper Agreement Limit=0.24 log(mm^3), and the Lower Agreement Limit=-0.19 log(mm^3). For an average volume of 1,000 mm^3, the agreement limits are 646 mm^3 and 1,738 mm^3.

Results across the participating groups are presented in Table 2. The analyses used to compute the overall statistics were conducted on the log-transformed scale. Stratified analyses were used to compute the repeatability metrics based on the original scale, for the small (< 4189mm^3 or a diameter less than about 20mm for a sphere) vs. large (> 4189mm^3 or diameter greater than about 20mm for a sphere) tumors. Seven of the 31 tumors were small by this criterion.

Table 2: Repeatability results: reporting CCC (Concordance Correlation Coefficient), RC (repeatability coefficient), RCLB and RCUB (lower bound and upper bound), and wCV (within-subject coefficient of variance).

	All Tumors			Small Tumors, vol<4189 mm^3 (n=7)			Large Tumors, vol>4189 mm^3 (n=24)		
Group	**CCC**	**RC [LB,UB]** (log)	**wCV** (%)	**CCC**	**RC [LB,UB]** (mm^3)	**wCV** (%)	**CCC**	**RC [LB,UB]** (mm^3)	**wCV** (%)
Group02	0.97	0.35 [.28,.47]	13.0	0.99	273 [184,522]	6.6	0.99	8913 [6927,12502]	11.3
Group03	0.71	1.5 [1.2,1.99]	54.0	0.77	1974 [1458,3055]	41.6	0.87	19170 [14277,29176]	36.1
Group04	1.00	0.06 [.05,.08]	2.1	0.99	326 [224,594]	6.5	1.00	2163 [1673,3061]	2.4
Group05	1.00	0.06 [.05,.08]	2.2	0.97	506 [334,1029]	10.3	1.00	3479 [2704,4881]	3.2
Group06	1.00	0.09 [.07,.12]	3.1	0.99	299 [206,546]	6.8	1.00	4117 [3184,5827]	4.9
Group07	1.00	0.09 [.07,.11]	3.1	0.98	390 [268,712]	8.9	1.00	4208 [3254,5955]	5.3
Group08	1.00	0.06 [.05,.08]	2.1	1.00	273 [188,499]	5.5	1.00	3536 [2734,5004]	4.1
Group11	0.98	0.36 [.29,.48]	13.0	0.64	1920 [1320,3505]	58.5	0.84	49929 [38615,70667]	46.4
Group12	0.99	0.22 [.17,.29]	7.8	0.90	1116 [768,2037]	25.0	0.91	51977 [40198,73565]	38.9
Group14	1.00	0.09 [.07,.12]	3.2	0.99	317 [214,606]	7.6	0.98	13069 [10158,18333]	13.3
Group15	1.00	0.11 [.08,.14]	3.8	0.98	567 [396,995]	12.3	1.00	2719 [2092,3886]	3.2
Group16	1.00	0.1 [.08,.13]	3.4	0.97	452 [299,920]	12.5	0.99	8682 [6779,12079]	10.2

1.3 Reproducibility Across Participating Groups

Reproducibility was assessed for various combinations of the tumors, as shown in Table 3.

Table 3: Number of tumors analyzed in each strata

Analysis	Strata	N
Overall	All	31
	Small	7
	Large	24
Profile=Yes	All	20
	Small	7
	Large	13
Profile=No	All	11
	Small	0
	Large	11
Human Edited	All	31
	Small	8
	Large	23
No editing	All	31
	Small	8
	Large	23

1.3.1 Determine Data to be Pooled

The individual repeatability results were inspected visually for bias and heteroscedasticity, and the data for those groups with similar distributions were pooled for the reproducibility analysis. Group03 was determined to have sufficiently different results than the other groups and was not included within the analysis. Additionally, review of the data exposed five anomalous readings (from three subjects by four groups). Volumes differed by log-orders of magnitudes from the rest of the data, suggesting data transcription errors. These were removed from the reproducibility analyses.

1.3.2 Analysis Across All Pooled Data

The pairwise performance across acquisition repetitions is illustrated in Figure 4. Like Figure 3, the left panel presents the data in the original scale of measurement (mm^3), while the Bland-Altman analysis was performed on log-transformed data due to the fact that the variation in measurement increased with volume. The 95% limits of agreement on the log10 scale are -0.35 and 0.39. Consider, for example, a mean volume reading of 1000 mm^3 (or 3 on the log10 scale). The Bland-Altman limits of agreement would be between 447 and 2455.

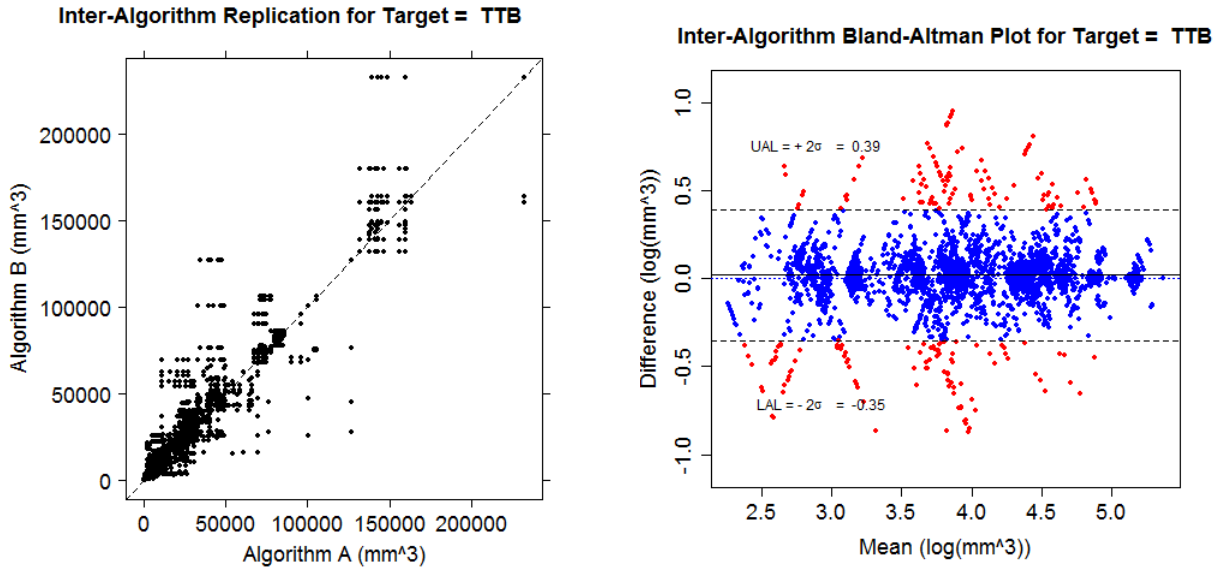


Figure 4: Results of Inter-algorithm analysis: Left panel: Scatter plot for all pairs of readers in the original scale of measurement. Data were log-transformed prior to Bland-Altman analysis. Right panel: Bland-Altman chart on the log-transformed data. The Upper Agreement Limit=0.39 log(mm^3), and the Lower Agreement Limit=-0.35 log(mm^3). For an average volume of 1,000 mm^3, the agreement limits are 447 mm^3 and 2455 mm^3.

The RDC was 0.37, which follows closely with the Bland and Altman limits. This value implies that we expect the difference between any two measurements taken on a subject regardless of reader or repetition is expected to be within +/- 0.37 log units, or about 14%.

Results of the Linear Mixed Effects (LME) model assess the degree to which algorithms contributed to overall variability versus that due to the subject or residuals. Model output is presented in Figure 6, depicted in a chart illustrating the weights of the four different variables on overall volume variability captured by the model.

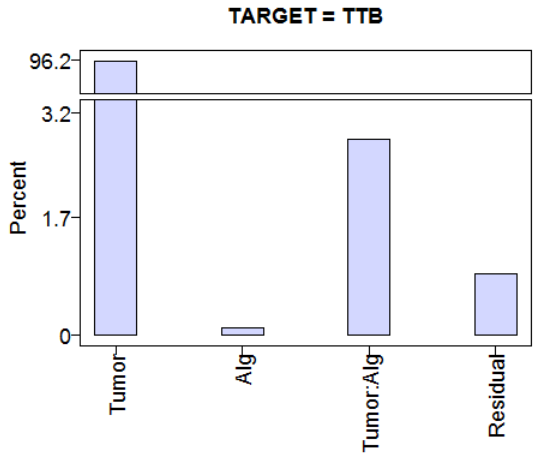


Figure 5: Results of LME for overall reproducibility analysis, represented in a Pareto chart of effect sizes

Tumor variation itself dominates with 96% of total variation, which is expected for meaningful biomarkers, in that this is the component which is due to the measurement itself. Tumor-by-algorithm interaction variance comprises the next highest variance, accounting for 3% of the variance, indicating that tumors are read differently by different algorithms, which is the primary reproducibility result. Residual variance of 1% accounted for factors including the test-retest variability itself which is not attributable to the algorithm performance.

Figure 6 shows the observed data (in mm^3 and $log(mm^3)$) plotted against fitted data (fitted data being an indication of how the model interpreted the data).

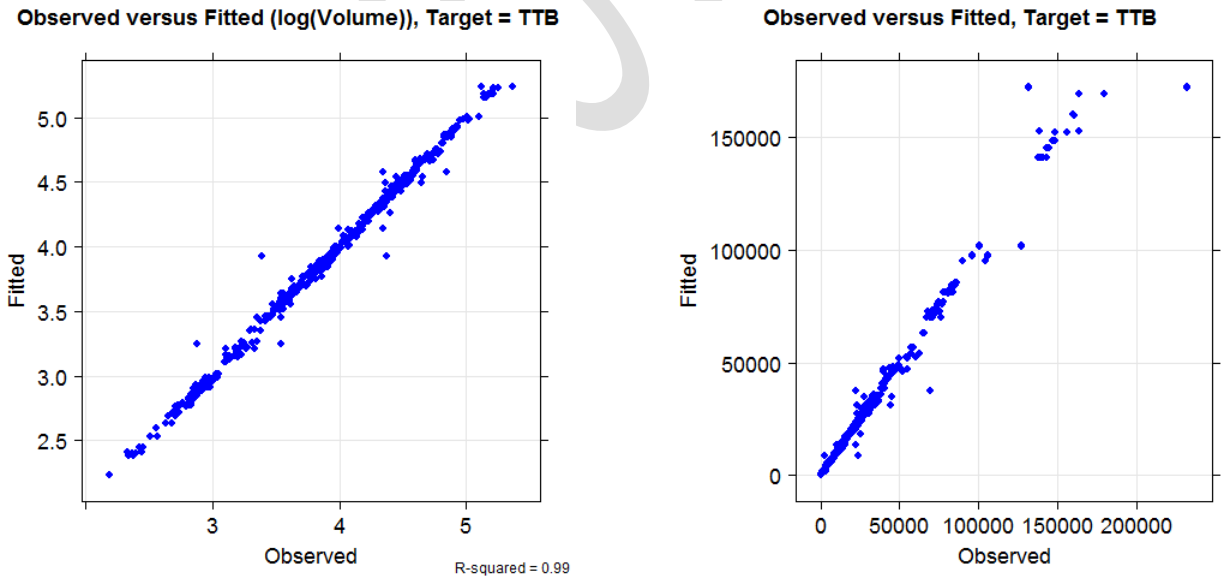


Figure 6: Observed vs. fitted on the log scale (left), and original scale (right).

Finally the Q-Q plot in Figure 7 demonstrates the linearity of a comparison of the distribution of residual volume values with a standard normal distribution. The Q-Q plot from this model

indicates that the core of the residuals follow a normal distribution. However, a handful of large residuals cause deviations from the expected distribution.

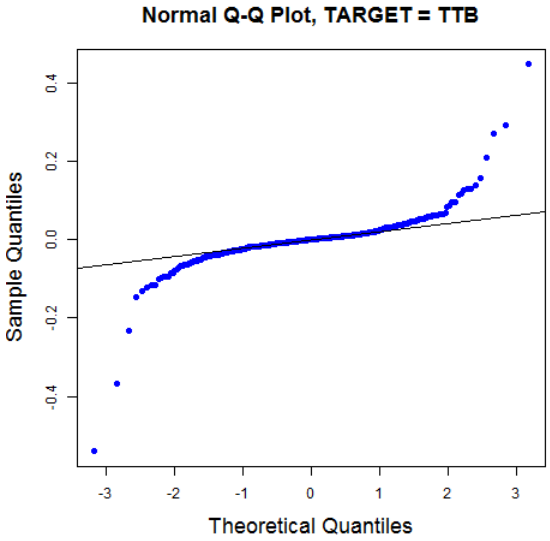


Figure 7: Q-Q plot indicating that the core of the residuals follow a normal distribution, but the distribution has long tails due to a handful of extreme values.

As indicated, the primary reproducibility due to algorithm is captured in the subject by reader interaction, as this quantifies the extent to which different algorithms measure different tumors differently. Figure 8 presents an interaction plot, where the x-axis = subject number ordered by average volume, and the y-axis = log(volume), with points colored and connected by algorithm. We see that the lines all trend together, but the lines cross each other. The crossing indicates that there is an interaction between subject and algorithm. (Alternatively, if all lines were parallel, then the interaction would be near-zero.)

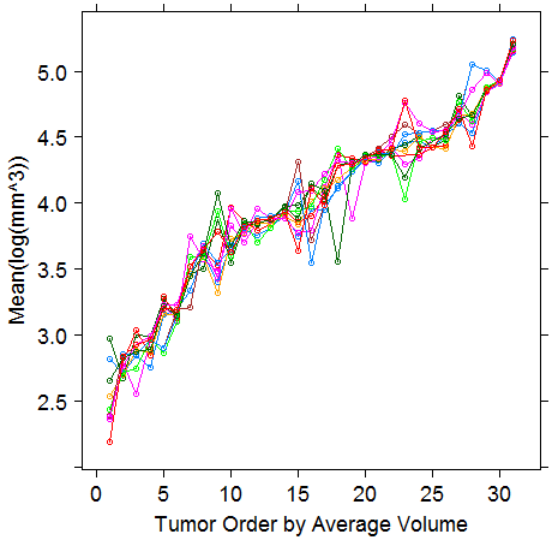


Figure 8: Parallel line plot. Tumors are ordered by average recorded volume across algorithms and test-retest values (x-axis). Points are connected by algorithm. The interaction term in the model accounts for the degree to which the lines cross in the figure.

1.3.3 Other Stratified Analyses

In addition to computing the metrics on all tumors, two stratified reproducibility analyses were performed. We looked at the effect on variability of the degree of automation used by the algorithm. We also looked at the effect on variability of classification of tumors, dividing them into two types: (a) tumors that could be classified as meeting the conditions described in the "Claims" section of the QIBA Profile, and (b) tumors that did not meet these conditions.

Four other stratified analyses were carried out similarly to that for all of the pooled data, as outlined in Table 3. Results for all 5 analyses are summarized in Table 4 below.

Table 4: Summary of Reproducibility Results to Inform QIBA Claim

	All Tumors					Small	Large
	95% LOA (log(mm^3))	**95% LOA (mm^3)**	**ICC**	**RDC (log)**	**Alg/Residual Variance (%:%)**	**RDC (mm^3)**	**RDC (mm^3)**
Combined	-.035/0.39	447/2455	0.96	0.37	3:1	1290	28205
Profile=Yes	-0.30/0.33	468/2138	0.97	0.32	2:1	1290	6369
Profile=No	-0.43/0.49	372/3090	0.87	0.45	10:2		41074
Editing	-0.38/0.39	417/2455	0.96	0.39	4:1	1343	26760
No editing	-0.25/0.38	562/2399	0.97	0.33	2:1	1234	33004

2.1 Create Reference Truth Segmentations

Figure 9 shows an example of a reference segmentation. One such reference segmentation was created for each test-retest repetition.

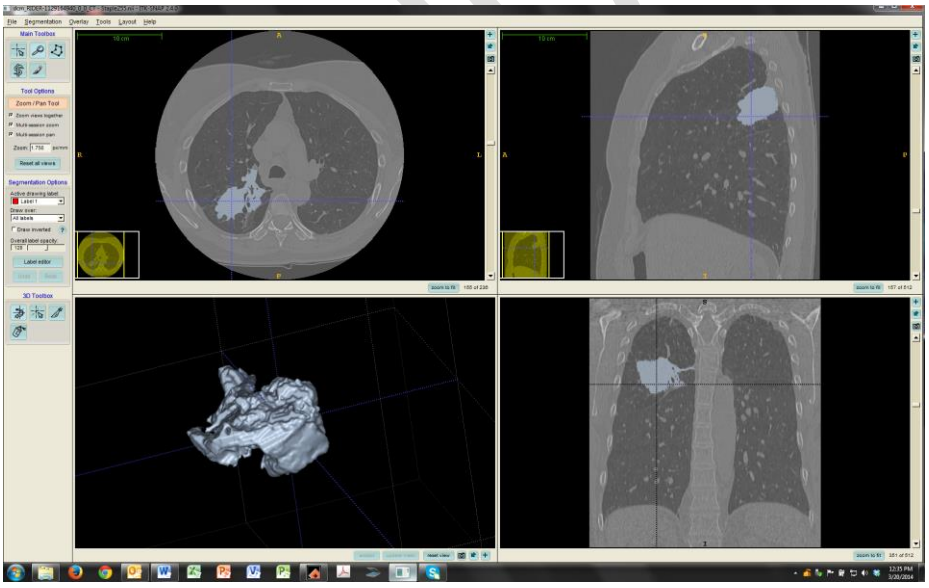


Figure 9: Example of a reference truth segmentation. (RIDER-1129164940, first repetition)

As indicated in the methods section, the reference segmentations were formed using an expectation-maximization algorithm. As a practical matter, that algorithm attempts to use all

input segmentations to influence the results according to the level of overlap among them. However, if one of the inputs is so far from the others so as to constitute a highly dissimilar input, the algorithm fails to produce a result. We had four such cases among our 62 cases (31 subjects x 2 test-retest repetitions), as summarized in Table 5.

Table 5: Cases where reference segmentation was generated by removing an outlier

Subject	Repetition	Outlier Group
2283289288	0	Group03
2283289288	1	Group03
1500037140	1	Group03
344011628	1	Group10

Additionally, reader segmentations from the QIBA 1B study were also utilized to create a separate reference used to represent typical results of readers on these same data.

Figure 10 shows an example of one participants' overlap with the corresponding reference segmentation. Of course, each participant can be overlaid accordingly and it is on this basis that the overlap metrics are computed.

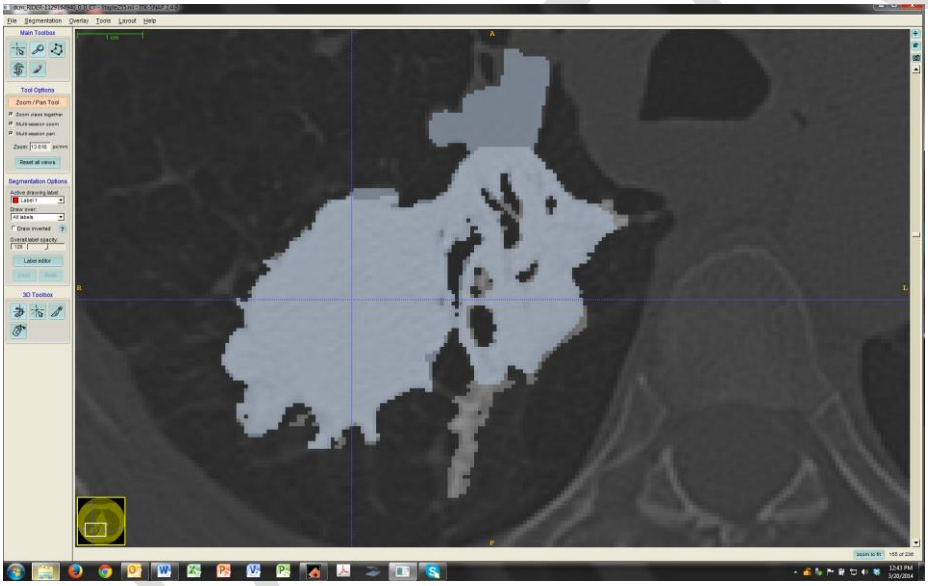


Figure 10: Example of a participating group's result superimposed on to the reference. TP voxels are rendered as light cyan, FN voxels as dark cyan, and FP as grey. TN pixels are displayed as reduced intensity background image. (RIDER-1129164940, first repetition, Group08)

2.3 Merging and Plotting of Histograms by Metric and Participant

Overlap metrics were calculated for each participant and test-retest repetition. A histogram of these results was created for each participant and merged onto a plot that compares the relative performance of each. The plots are shown in Figure 11.

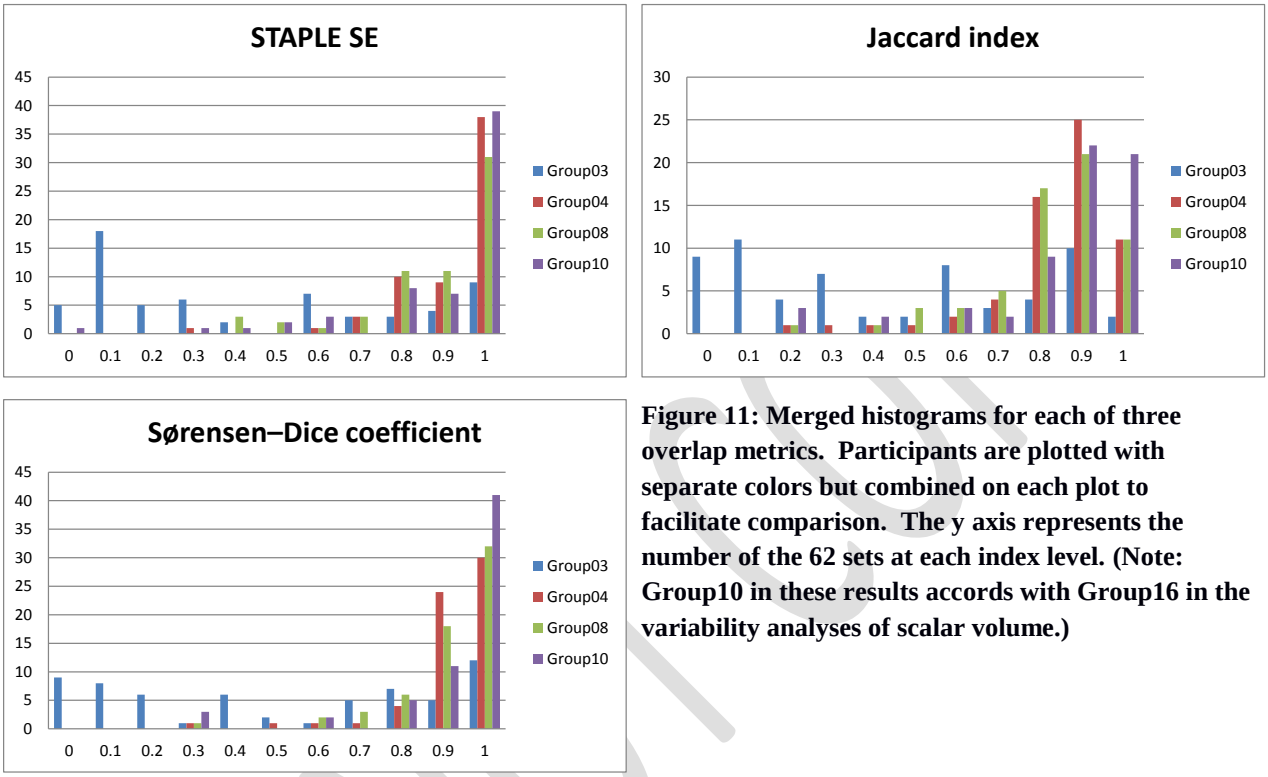


Figure 11: Merged histograms for each of three overlap metrics. Participants are plotted with separate colors but combined on each plot to facilitate comparison. The y axis represents the number of the 62 sets at each index level. (Note: Group10 in these results accords with Group16 in the variability analyses of scalar volume.)

IV. DISCUSSION

Analysis of test-retest repeatability provides metrics about what level of repeatability to expect from individual algorithms. Overall, repeatability was high as indicated by CCC near 1 and wCVs below 10% for most participating groups. Group03 performed the least well, Groups 02, 11, and 12 moderately well, and other groups better. Individualized reports inclusive of raw data and intermediate analysis results have been provided to participating groups in the challenge. The value of the results is highest to those who contributed actual segmentation boundaries. The results of analyzing reproducibility across the algorithms are summarized in Table 4, and can be used to validate the QIBA claim. The RDC values are particularly instructive. The RDC values approximately correspond with that participant with the least performing repeatability value (plus a fraction) (see Table 2).

By these results, the best case is when algorithms are used on tumors meeting the measurability criterion defined in the Profile, and the worst performance for tumors not meeting measurability criteria, with variability being approximately 1-1/2 times as much, by a comparison of RDC values of .32 versus .45, respectively. The value of editing was equivocal; smaller tumors appear to be better without editing, but larger tumors benefit by editing. This may be intuitive, in that larger tumors more often include features which may or may not be considered tumor mass.

An additional consideration for which these data are informative concerns the extent to which the algorithm may be considered "the end of the line" with respect to variability of the entire process of evaluating tumor size. One line of thinking is that the algorithm's performance dominates variability earlier in the processing chain and that compliance may be given by performing at or better than the overall system claim. These data suggest that there is as much as one-half of the variability coming from sources independent of the algorithms, to as little as one-fifth. On this basis, no more than two-thirds of the overall variability claim of the system as a whole can be allocated to analysis software if the overall system is to be compliant (or less if the scanner is not held constant).

The greatest utility of this work from a participant's point of view, or a company seeking to commercialize analysis software for tumor volumetry, is a comparison of their algorithm with other similar algorithms' performance, and the measure of a performance standard that can be defined by QIBA by this type of analysis. Participants also benefit by algorithm comparisons to identify weaknesses of their algorithms and areas needing improvement. This is greatly aided by the results of the segmentation object analysis, which provides insight into why volume calculations under- or over-estimate a volume (see Figure 10 for examples of this). Some particularly illustrative cases, each representing different circumstances, are shown in Table 6. Full evaluation of these results is beyond the scope of the present study but the detailed maps are provided to participants who contributed segmentation objects.

Table 6: Interesting cases

SUBJID	REP	SE Group 03	SE Group 04	SE Group 08	SE Group 10	3A Volume (mean, stdev)	1B Volume (mean, stdev)	Notes
1129164940	0	0.55	0.94	0.94	0.86	49cc, 24cc	44cc,14cc	Seems typical
1500037140	0	0.91	0.92	0.83	0.97	7cc,.1.4cc	6cc,0.5cc	Seems typical
1760553574	0	0.04	0.96	0.77	0.86	9cc,12cc	3cc,9cc	All struggled
1801720707	1	0.22	0.87	0.37	0.30	45cc,289cc	0.6cc,0.6cc	Grp 5 and 12 very high
2016615262	0	0.00	0.93	0.89	0.97	25cc,23cc	20cc,4cc	Grp 3 low, Grp 11 high
2151469008	0	0.13	0.92	0.91	0.91	29cc,16cc	27cc,22cc	Algs tighter here (except Grp 3)
2357766186	0	0.25	0.85	0.96	0.90	16cc,19cc	2cc,4cc	Grp 3 and 15 like rdrs
2539508879	0	0.86	0.95	0.92	0.91	8cc,1cc	7cc,3cc	Algs tighter here
2619750334	1	0.58	0.90	0.94	0.97	80cc,49cc	69cc,59cc	Algs tighter here (except Grp 3)
2799584460	1	0.97	0.68	0.71	0.97	0.6cc,0.3cc	0.8cc,0.5cc	Algs tighter here
3115188676	0	0.00	0.77	0.99	0.72	11cc,9cc	12cc,4cc	Grp 3 much lower
5195703382	1	0.14	0.27	0.70	0.73	27cc,40cc	0.6cc,0.3cc	Grp 15 like rdrs, rest different

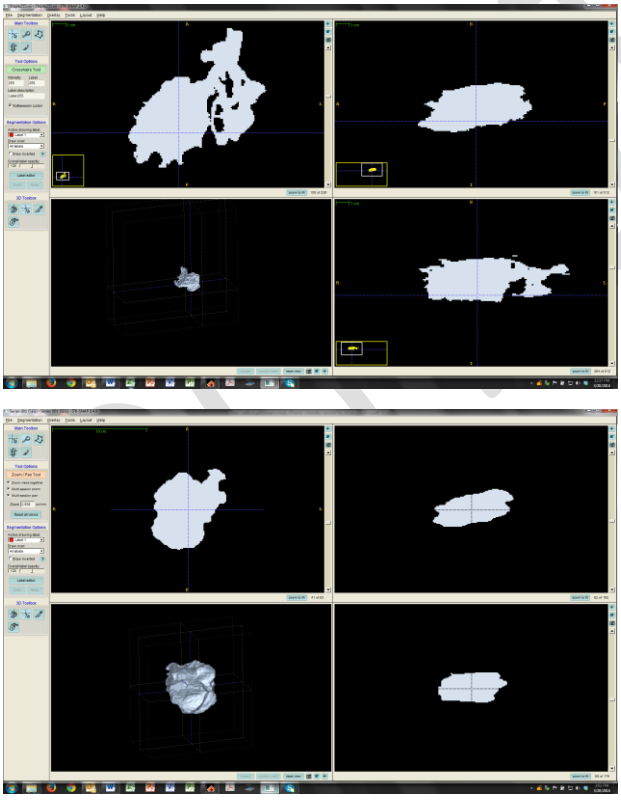


Figure 12: Comparison of algorithm derived reference versus manual contouring derived reference. Note that in general the algorithm result has more jagged edges and includes anatomic features not included by readers. In this example, the average volume resulting from algorithms was 49,435mm^3 whereas readers 43,731mm^3. In general algorithm segmentation result in larger volumes though the factor varies by degree of

subject difficulty. (Positioned to the same slice but scaling differs based on technical reasons not relevant to the comparison.) (RIDER-1129164940, first repetition)

Finally, Figure 13 shows a visual comparison of the performance of the 12 participating groups.

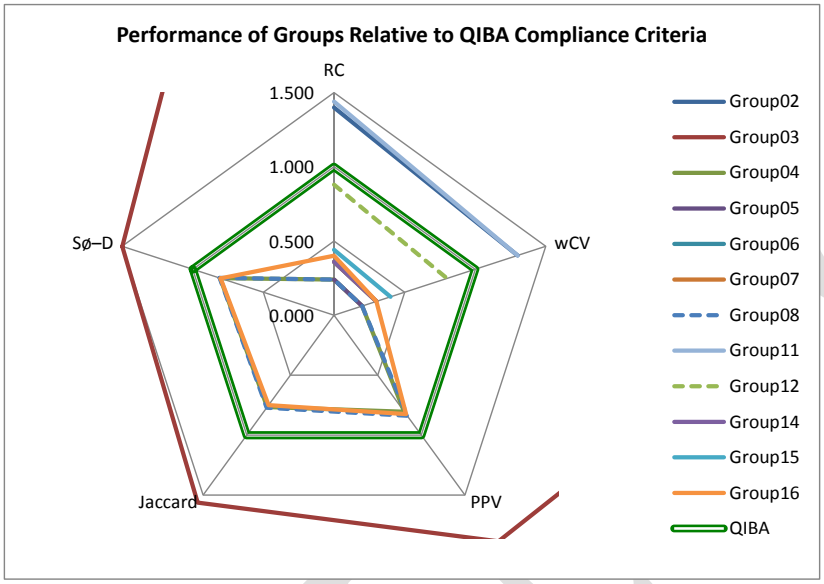


Figure 13: Aggregate chart proposing a way to represent the relative performance of groups and how they relate to the QIBA Claim

RC tells cross-sectional variability as an absolute measure, wCV provides a weighted indication, and the overlap metrics indicate the degree to which the volumes were computed on an appropriate segmentation.

V. CONCLUSIONS

Based on the specific role of tumor volumetry in clinical practice, we have computed performance metrics critical to the role of the biomarker including repeatability and reducibility of scalar volumes as well as overlap measures computed from analyzing segmentation objects favoring consistency. For measurement of tumor volume to be used as a predictor of true biological feature change or difference, tumor volume must predictably reflect the true and biologically-relevant feature measurement, dependent upon results having a high standard for repeatability and reproducibility. In addition to repeatability and reproducibility variability, this study contributes to the understanding of whether resulting segmentations reliably represent what would be considered the actual segmentation of the given tumor. We have described methods for estimating these metrics, and applied them to twelve specific algorithms on non-small cell lung cancer datasets. Going forward, these metrics can be computed on larger reference data sets representing explicitly described sub-populations, e.g., as defined in the QIBA Profile clinical context for use statement. Additionally, the procedure used here may be suited for use as a means by which compliance of analysis software may be rigorously determined and reported.

Our study has limitations. One stems from the fact that definitions of fully- versus semi-automated algorithm processing evolved during the course of the study and as such more rigorous investigation of differing categories have been suggested. Another limitation stems from an explicit determination for this study that workflow not be constrained, but the related 1B study suggests its importance. We had determined that automatic algorithms would not differ in their performance based on workflow, but found that this does not always hold true. Additionally, the data used in this study were relatively limited. Although the data was contained an assortment of clinical cases, it did not fully represent the claimed clinical context of use for the corresponding QIBA Profile. Definitive reference data sets that adequately represent the target patient population according to formally assessed statistical criteria should include patients representing a range of common co-morbidities, disease characteristics, and imaging settings (e.g. sedated vs. non-sedated patients). Finally, the manner in which these tests are run and the data collected has implications regarding the interpretation and use of metrics computed and reported. For example, execution of these tests by a trusted third-party on sequestered data sets may increase their utility.

VI. ACKNOWLEDGEMENTS

The challenge study could not have taken place without the active participation by the organizations who submitted data. Specifically, we acknowledge the personal attention by the following:

LIST THE SPECIFIC POINTS OF CONTACT AT THE PARTIPANTS

We also acknowledge the crucial logistical support from the RSNA staff in administering the application process which included anonymized interactions among study participants, funding provided by NIBIB to defray statistical analysis costs, and of course the participating groups themselves as without their effort to produce the submissions there would have been no project. We also acknowledge Mike McNitt-Gray of UCLA for providing the tumor measurability strata from the 1B project which he led, as well as David Clunie who provided segmentation objects from the readers in the 1B study.

VII. APPENDIX: ALGORITHM DESCRIPTIONS

Twelve groups participated in the challenge by submitting volume readings and five submitted segmentation objects, four of which were compatible for analysis. Algorithms from each group are described below.

Participant	Description / Workflow
Group02 (volume readings and segmentation objects[1]) Moderate image/boundary modification (on less than 50% of the tumors)	Volumetric analysis was determined using a segmentation approach employing a Z-score on the highest conspicuity post-contrast volumetric image set. A cylinder is placed around the highest conspicuity slice and around all slices above and below this slice in which the tumor is seen. A kernel defined within the region of interest (ROI) is then propagated to other slices using connectivity algorithms. The search is constrained by the predefined cylinder to accelerate the search algorithm.
Group03 (volume readings and segmentation objects) Fully automatic	One-click user-seeded segmentation. Utilizes shape and boundary information to delineate the tumor. The workflow for segmenting lung tumors involves a single click at a seed-point roughly centered in the tumor. The algorithm uses the seed point in combination with a thresholded ROI in order to extract the most probable shape of the tumor.
Group04 (volume readings and segmentation objects) Automated and semi-automated: limited image/boundary modification (on less than 15% of the tumors)	Utilize a trained non-radiologist technician and trained radiologist. As the images would be of chest and the tumors would be in lung parenchyma, all the volume assessment were made using a fixed lung window/level display setting of 200HU (window) and -1400HU (level). Trained non-radiologist opens the images in and uses the tumor location to identify the tumors on images. Trained non-radiologist outlines/ROIs of the identified tumors using automated segmentation tools. Trained non-radiologist evaluates the quality of the segmentation and adjusts outlines with additional semi-automated tools as necessary. Finally, that image data is submitted to trained radiologist for final assessment of outlines/ROIs. The trained radiologist evaluates the quality of the segmentation and adjusts outlines with automated & semi-automated tools as necessary. Once trained radiologist is satisfied with all the outlines/ROIs of the respective tumors, the automated volume assessment tool is used to calculate volume as volume = (Image Position Interval1 * Area1) + (Image Position Interval2 * Area2)...+...+ (Image Position Interval n * Area n). The images with ROI is processed, re-colored and converted in to .nii file.
Group05 (volume readings) Semi-automatic; Moderate parameter adjustment (on less than 50% of the tumors).	Modelization of the heat-flow between the inside and outside of the tumor. Based on intensity gradients, in 3D. User clicks on a tumor, or draws a diameter joining the boundaries of the tumor => software computes a segmentation of the tumor, and displays its contours. User can then refine the segmentation by the means of a slider => software adjusts the segmentation accordingly, and displays in real-time the new contours. If needed, user can manually edit any contour by drawing it. User finally validates the segmentation => software “locks” the segmentation and extracts the statistics: volume, long axis, short axis, and all intensity-based numbers (average value, standard deviation, etc.)
Group06 (volume readings) Fully automatic; (uses only seed points and ROI information)	This algorithm combines the image analysis techniques of region-based active contours and level set approach in a unique way to measure tumor volumes. It may also detect volume changes in part solid and Ground Glass Opacity tumors. The user clicks and drags to define an elliptical/circle ROI to initiate the segmentation.

[1] Alignment issues prevented inclusion in the segmentation object analysis.

Participant	Description / Workflow
	The computer then carries out the segmentation, and tumor measurements are saved. The algorithm is an edge-based segmentation method that uniquely combines the image processing techniques of marker-controlled watershed and active contours. An operator initializes the algorithm by manually drawing a region-of-interest encompassing the tumor on a single slice and then the watershed method generates an initial surface of the tumor in three dimensions, which is refined by the active contours. The volume, maximum diameter and maximum perpendicular diameter of a segmented tumor are then calculated automatically.
Group07 (volume readings) Fully automatic; (uses only seed points and ROI information)	An initialization sphere is drawn from the center of the mass, on the slice with its largest extents, such that it covers the entire extent of the mass. The user determines the center and radius in a single click-drag action, and this initialization circle imposes hard constraints on the maximum extents of the three dimensional segmentation. The employed segmentation tool is part of a commercial software package for multimodal oncology treatment assessment and review. Thus the workflow mimics the typical workflow a user has with this tool: Select the desired CT data set and load it into any review mode Select the lung window-level setting Navigate to the tumor center using the pixel and slice locations from the MSKCC Coffee Break study Locate the slice where the tumor has the greatest extents Select the segmentation tool, and initialize the segmentation by clicking in the approximate center of the mass and dragging the mouse to set the radius of the spherical region of interest. The spherical region of interest contains a fixed inner sphere and the outside sphere which is set by the mouse dragging motion. The radius is chosen such that the inner circle encompasses most of the mass to be segmented, and the outer sphere can be used as a constraint to prevent any leakage into the chest wall or heart if the mass is attached/abducting to these organs. The computation takes a few seconds (single digit numbers) to compute the result. User may retry the segmentation a few times if the result is unsatisfactory. With each try the previous result is erased, and does not influence the result of preceding try. In this experiment, the user has in overall three tries to get a satisfactorily result. Once the segmentation has been determined, the user reads off the volume from the region statistics, which are automatically computed and displayed as soon as the segmentation has been defined. (The volume estimation algorithm counts all voxels whose centroid lies within the segmented contour and multiplies this number with voxel volume) To document the segmentation result, save the segmentation as a RT-structure set to the data repository.
Group08 (volume readings and segmentation objects) Semi-automatic; Moderate parameter adjustment (on less than 50% of the tumors)	Semi-automatic segmentation based on thresholds, growing region and mathematical morphology processing DICOM images are downloaded and imported into a database. Image data are converted to a proprietary optimized format before the insertion into the database. Tumors coordinate are downloaded and reformatted by our data manager. Relying on a proprietary Validation Framework System, landmarks are automatically inserted into the database. The software is allowed then to display the repeated images side by side with the correct landmarks identifying the tumors to segment. The first repetition was edited as a single image. The side-by-side displayed was available only for the repetition when the first scan edit was locked. Three reviewers are involved, each in charge of segmenting approximately a third of the dataset. The data manager made available to the reviewers a commercial semi-automated segmentation tool dedicated to Lung tumors. Another manual tool can be enabled if semi-automatic segmentations were not fully satisfactory. The data manager recommended using different window level to better assess tumors boundary, pulmonary window level being the major window level to refer to. The data manager recommended correcting semi-automated segmentation as long as the segmentation was not fully satisfactory. Once the whole dataset segmented, an additional reviewer was involved to check the whole coherency of the measurements: Total number of tumors, no obvious incoherency, correct recording of the data, etc. A complete report was extracted. The same Validation Framework System allowed automatic extraction of tumors mask as .mhd format. A third party software as SLICER was used to convert masks to NIFTI format.
Group11 (volume readings) Fully automatic (uses only seed points and ROI information)	Method is completely automatic and consists of three steps. First, a region of interest is extracted and the tumor is classified as solid or subsolid. In the second step, a binary segmentation mask is computed by an algorithm based on thresholding and morphological postprocessing, using slightly different procedures for the two classes. Finally, the volume of the tumor is determined by adaptive volume averaging correction.

Participant	Description / Workflow
	Preprocessing: a stroke is generated from the given center and bounding box by shortening the bounding box diameter to 40%. The segmentation is performed in a cubic region of interest (ROI), whose edge length is twice the stroke length. The ROI is smoothed with a 3 x 3 Gaussian filter and resampled to isotropic voxels and a maximum size of 100 x 100 x 100 voxels. For detecting the tumor type, the local maximum in a 5 x 5 x 5 neighborhood of the ROI center is identified. If its value is greater than -475 HU, the tumor is treated as solid, otherwise as subsolid. The ROI center is used as a seed point for region growing. The lower threshold is derived from the 55% quantile of the histogram of the dilated stroke by applying an optimal elliptic function yielding values between -780 and -450 HU. The resulting mask contains the complete tumor, but may also leak into adjacent vasculature or, in case of juxtapleural tumors, into structures outside the lungs. In order to remove vessels, an adaptive opening is applied, where the erosion threshold is chosen such that the segmentation has no connection to the ROI boundary anymore. A slight overdilation allows a final refinement of the mask. In order to avoid leakage outside the lungs, a convex hull of the lung parenchyma is computed within a minimal elliptical region that is fitted to the shape of the tumor. The convex hull is then used as a blocker for the segmentation. Due to the limited spatial resolution of CT and partial volume effects, the volume of a segmented tumor cannot be determined exactly by voxel counting. Instead, voxels in a tube around the segmentation boundary are weighted according to their estimated contribution to the tumor volume. The weight depends on the relation of a voxel's value to the typical tumor and parenchyma densities.
Group12 (volume readings) One with interactive correction - moderate image/ boundary modification (on less than 50% of the tumors)	We start with an automatic method (submitted Group11) and correct results interactively if necessary. The user draws partial contours which are included in the segmentation in the edited slice. Additionally, the correction is automatically propagated to a set of neighboring slices by sampling the contour, matching points to the next slice and connecting them with a live-wire method. Interactive correction: Our interactive correction tool provides an efficient way to fix segmentation results which are mostly correct but need some refinement. The user draws partial contours indicating the desired segmentations which are then automatically propagated into 3d. Seed points calculated from the user contour are moved to adjacent slices by a block matching algorithm and the seed points are connected by a live-wire algorithm. Details can be found in our paper {reference provided separately}. For the submission, correction was performed by two experienced developers in consensus. Volumetry: The volumetry used for automatic results is integrated in the segmentation algorithm. To ensure consistency after interactive correction, the change in the number of voxels is computed and multiplied with the (partial-volume-corrected) volume of the initial result.
Group14 (volume readings) Fully automatic (uses only seed points and ROI information)	The system is fully automated after manual input of an approximate bounding box for the tumor of interest. Within the bounding box, the system automatically processes the images in 3 stages-preprocessing, initial segmentation, and 3D level-set segmentation. In the first stage, a set of smoothed images and a set of gradient images are obtained by applying 3D preprocessing techniques to the original CT images. Smoothing, anisotropic diffusion, gradient filtering, and rank transform of the gradient magnitude are used to obtain a set of edge images. In the second stage, based on attenuation, gradient, and location, a subset of pixels is selected, which are relatively close to the center of the tumor and belong to smooth (low gradient) areas. The pixels are selected within an ellipsoid that has axis lengths one-half of those of the inscribed ellipsoid within the bounding box. This subset of pixels is considered to be a statistical sample of the full population of pixels in the tumor. The mean and SD of the intensity values of the pixels belonging to the subset are calculated. The preliminary tumor contour is obtained after thresholding and includes the set of pixels falling within 3 SDs of the mean and with values above the fixed background threshold. A morphologic dilation filter, a 3D flood fill algorithm, and a morphologic erosion filter are applied to the contour to connect the nearby components and extract an initial segmentation surface. The size of the ellipsoid and the remaining parameters are selected experimentally in a way that enables segmentation of a variety of tumors, including necrotic tumors. In the third stage, the initial segmentation surface is propagated by using a 3D level-set method. Four level sets are applied sequentially to the initial contour. The first three level sets are applied in 3D with a predefined schedule of parameters, and the last level set is applied in 2D to every section of the resulting 3D segmentation to obtain the final contour. The first level set slightly expands and smooths the initial contour. The second level set pulls the contour toward the sharp edges, but at the same time, it expands slightly in regions of low gradient. The third level set further draws the contour toward the sharp edges. The 2D level set performs final refinement of the segmented contour on every section.
Group15 (volume readings) Moderate image/boundary modification (on less than 50% of	The software used is essentially a semi-automated contouring method. The user clicks on a voxel located inside the tumor of interest and then drag a line to the outside of the tumor (to the background). The voxels along that line are sampled and a histogram of intensities (Hounsfield Units) is created.

Participant	Description / Workflow
the tumors)	A statistical method is employed to determine the threshold that best separates the two distributions (tumor and background) in that histogram. Once that threshold is determined, the software employs a 3-D (or if selected a 2-D) seeded region growing using the initial voxel selected as the point inside the tumor and the threshold determined from the histogram analysis. The tool also provides several user editing tools such as adding and erasing voxels from the contour, etc. The workflow description: Each contour is automatically stored in a database linked to the experiment along with meta data such as patient id, contouring individual's id, etc. Each contoured object has a unique id that is linked to the series uid to maintain its identity. Once the contour is completed and accepted, the volume of the contoured object is calculated. This is done essentially by counting the number of voxels within the boundaries of the contoured object and multiplying that by the voxel size (as derived from DICOM header data).
Group16 (volume readings and segmentation objects[2]) Limited image; boundary modification (on less than 50% of the tumors)	As the input for the algorithm, the user has to draw a stroke being favorably the largest diameter in the axial orientation or click a point in the given lung tumor. Usually, the decision to use a stroke or a single click point depends on the size of the tumor to be segmented (for bigger tumors, a stroke is preferable, while for small tumors, a single click is sufficient). In the next step, a Volume of Interest (VOI) around the tumor is estimated. In the case where the algorithm has been initialized with stroke, the size of the VOI depends on the length of the stroke. 3D region growing is conducted in a VOI starting from seeds generated along the stroke or around the click point, depending on the initialization. Adjacent structures of similar density (pleura, vessels) are separated by a set of interchanging morphological operations (erosion, dilation, convex hull and binary combination with region growing mask.) Finally, a plausibility check between the resulting segmentation mask and the position of the initial stroke or click point is conducted. If necessary, initial thresholds are re-adjusted and the whole procedure (steps 2-5) is repeated. For the case when the semi-automatic results are not satisfactory, the software provides the possibility of correcting the results by drawing contours in selected slices and then propagating the contours in an automatic manner onto the whole 3D segmentation. The algorithm performs best optimally for the resolution up to 2 mm, though it still works reasonably well for thicker slices such as 5 mm.

Three groups (Group01, Group9, and Group13) initially applied but did not submit results.

[2] Segmentation objects submitted under Group10 ID.

VIII. REFERENCES

1. Biomarkers Definitions Working Group, *Biomarkers and surrgate endpoints: preferred definitions and conceptual framework.* Clinical Pharmacology and Therapeutics, 2001. **69**(3): p. 89-95.
2. Woodcock, J. and R. Woosley, *The FDA critical path initiative and its influence on new drug development.* Annual Review of Medicine, 2008. **59**: p. 1-12.
3. QIBA-Performance-Working-Group, *Review of Statistical Methods for Technical Performance Assessment.* Submitted to SMMR., 2014.
4. QIBA-Metrology-Metrology-Working-Group, *The Emerging Science of Quantitative Imaging Biomarkers: Terminology and Definitions for Scientific Studies and for Regulatory Submissions.* Submitted to SMMR., 2014.
5. Gurland, J. and R.O. Johnson, *Case for using only maximum diameter in measuring tumors.* Cancer Chemother Rep, 1966. **50**(3): p. 119-24.
6. Moertel, C.G. and J.A. Hanley, *The effect of measuring error on the results of therapeutic trials in advanced cancer.* Cancer, 1976. **38**(1): p. 388-94.
7. Royal, H.D., *Technology assessment: scientific challenges.* AJR Am J Roentgenol, 1994. **163**(3): p. 503-7.
8. Buckler, A.J., P.D. Mozley, L. Schwartz, N. Petrick, et al., *Volumetric CT in lung cancer: an example for the qualification of imaging as a biomarker.* Academic radiology, 2010. **17**(1): p. 107-15.
9. Mozley, P.D., L.H. Schwartz, C. Bendtsen, B. Zhao, N. Petrick, and A.J. Buckler, *Change in lung tumor volume as a biomarker of treatment response: a critical review of the evidence.* Ann Oncol, 2010. **21**(9): p. 1751-5.
10. Buckler, A.J., *A procedural template for the qualification of imaging as a biomarker, using volumetric CT as an example*, in *IEEE Applied Imagery Pattern Recognition Workshop*2009: Cosmos Club, Washington, D.C. p. 7.
11. Buckler, A.J., J.L. Mulshine, R. Gottlieb, B. Zhao, P.D. Mozley, and L. Schwartz, *The use of volumetric CT as an imaging biomarker in lung cancer.* Academic radiology, 2010. **17**(1): p. 100-6.
12. Buckler AJ, S.L., Petrick N, McNitt-Gray M, Zhao B, Fenimore C, Reeves AP, Mozley PD, Avila RS, *Data Sets for the Qualification of CT as a Quantitative Imaging Biomarker in Lung Cancer.* Optics express, 2010. **18**(14): p. 16.
13. Shankar, L.K., A. Van den Abbeele, J. Yap, R. Benjamin, S. Scheutze, and T.J. Fitzgerald, *Considerations for the use of imaging tools for phase II treatment trials in oncology.* Clinical cancer research : an official journal of the American Association for Cancer Research, 2009. **15**(6): p. 1891-7.
14. Zhao, B., L.H. Schwartz, and S.M. Larson, *Imaging surrogates of tumor response to therapy: anatomic and functional biomarkers.* J Nucl Med, 2009. **50**(2): p. 239-49.
15. Maitland, M.L., *Volumes to learn: advancing therapeutics with innovative computed tomography image data analysis.* Clin Cancer Res, 2010. **16**(18): p. 4493-5.
16. Nishino, M., D.M. Jackman, H. Hatabu, P.A. Janne, B.E. Johnson, and A.D. Van den Abbeele, *Imaging of lung cancer in the era of molecular medicine.* Acad Radiol, 2011. **18**(4): p. 424-36.

17. Koshariya, M., R.B. Jagad, J. Kawamoto, P. Papastratis, H. Kefalourous, T. Porfiris, C. Tzouma, and N.J. Lygidakis, *An update and our experience with metastatic liver disease.* Hepatogastroenterology, 2007. **54**(80): p. 2232-9.
18. Jaffe, C.C., *Measures of response: RECIST, WHO, and new alternatives.* J Clin Oncol, 2006. **24**(20): p. 3245-51.
19. Gavrielides, M.A., L.M. Kinnard, K.J. Myers, and N. Petrick, *Noncalcified lung nodules: volumetric assessment with thoracic CT.* Radiology, 2009. **251**(1): p. 26-37.
20. Morgensztern, D., S. Waqar, J. Subramanian, F. Gao, K. Trinkaus, and R. Govindan, *Prognostic Significance of Tumor Size in Patients with Stage III Non-Small-Cell Lung Cancer: A Surveillance, Epidemiology, and End Results (SEER) Survey from 1998 to 2003.* J Thorac Oncol, 2012. **7**(10): p. 1479-84.
21. Goldstraw, P., J. Crowley, K. Chansky, D.J. Giroux, P.A. Groome, R. Rami-Porta, P.E. Postmus, V. Rusch, and L. Sobin, *The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours.* J Thorac Oncol, 2007. **2**(8): p. 706-14.
22. Moskowitz, C.S., X. Jia, L.H. Schwartz, and M. Gonen, *A simulation study to evaluate the impact of the number of lesions measured on response assessment.* Eur J Cancer, 2009. **45**(2): p. 300-10.
23. Choi, H., *Response evaluation of gastrointestinal stromal tumors.* Oncologist, 2008. **13 Suppl 2**: p. 4-7.
24. Petrick, N., H.J.G. Kim, D. Clunie, K. Borradaile, et al., *Evaluation of 1D, 2D and 3D nodule size estimation by radiologists for spherical and non-spherical nodules through CT thoracic phantom imaging*, in *SPIE 2011*.
25. Kinnard, L.M., M.A. Gavrielides, K.J. Myers, R. Zeng, J. Peregoy, W. Pritchard, J.W. Karanian, and N. Petrick, *Volume error analysis for lung nodules attached to pulmonary vessels in an anthropomorphic thoracic phantom.* Proc SPIE, 2008. **6915**: p. 69152Q; doi:10.1117/12.773039.
26. Gavrielides, M.A., R. Zeng, L.M. Kinnard, K.J. Myers, and N. Petrick, *A template-based approach for the analysis of lung nodules in a volumetric CT phantom study.* Proc SPIE, 2009. **7260**: p. 726009; doi:10.1117/12.813560.
27. Winer-Muram, H.T., S.G. Jennings, C.A. Meyer, Y. Liang, A.M. Aisen, R.D. Tarver, and R.C. McGarry, *Effect of varying CT section width on volumetric measurement of lung tumors and application of compensatory equations.* Radiology, 2003. **229**(1): p. 184-94.
28. Ravenel, J.G., W.M. Leue, P.J. Nietert, J.V. Miller, K.K. Taylor, and G.A. Silvestri, *Pulmonary nodule volume: effects of reconstruction parameters on automated measurements--a phantom study.* Radiology, 2008. **247**(2): p. 400-8.
29. Borradaile, K. and R. Ford, *Discordance between BICR readers.* Appl Clin Trials, 2010. **Nov 1**: p. Epub.
30. Gavrielides, M.A., R. Zeng, K.J. Myers, B. Sahiner, and N. Petrick, *Benefit of Overlapping Reconstruction for Improving the Quantitative Assessment of CT Lung Nodule Volume.* Acad Radiol, 2012.
31. Gavrielides, M.A., R. Zeng, L.M. Kinnard, K.J. Myers, and N. Petrick, *Information-theoretic approach for analyzing bias and variance in lung nodule size estimation with CT: a phantom study.* IEEE Trans Med Imaging, 2010. **29**(10): p. 1795-807.
32. Gavrielides, M.A., L.M. Kinnard, K.J. Myers, J. Peregoy, W.F. Pritchard, R. Zeng, J. Esparza, J. Karanian, and N. Petrick, *A resource for the assessment of lung nodule size*

estimation methods: database of thoracic CT scans of an anthropomorphic phantom. Opt Express, 2010. **18**(14): p. 15244-55.
33. Das, M., J. Ley-Zaporozhan, H.A. Gietema, A. Czech, et al., *Accuracy of automated volumetry of pulmonary nodules across different multislice CT scanners.* Eur Radiol, 2007. **17**(8): p. 1979-84.
34. Bolte, H., C. Riedel, S. Muller-Hulsbeck, S. Freitag-Wolf, G. Kohl, T. Drews, M. Heller, and J. Biederer, *Precision of computer-aided volumetry of artificial small solid pulmonary nodules in ex vivo porcine lungs.* Br J Radiol, 2007. **80**(954): p. 414-21.
35. Cagnon, C.H., D.D. Cody, M.F. McNitt-Gray, J.A. Seibert, P.F. Judy, and D.R. Aberle, *Description and implementation of a quality control program in an imaging-based clinical trial.* Acad Radiol, 2006. **13**(11): p. 1431-41.
36. Goodsitt, M.M., H.P. Chan, T.W. Way, S.C. Larson, E.G. Christodoulou, and J. Kim, *Accuracy of the CT numbers of simulated lung nodules imaged with multi-detector CT scanners.* Med Phys, 2006. **33**(8): p. 3006-17.
37. Oda, S., K. Awai, K. Murao, A. Ozawa, Y. Yanaga, K. Kawanaka, and Y. Yamashita, *Computer-aided volumetry of pulmonary nodules exhibiting ground-glass opacity at MDCT.* AJR Am J Roentgenol, 2010. **194**(2): p. 398-406.
38. McNitt-Gray, M.F., L.M. Bidaut, S.G. Armato, C.R. Meyer, et al., *Computed tomography assessment of response to therapy: tumor volume change measurement, truth data, and error.* Transl Oncol, 2009. **2**(4): p. 216-22.
39. Keil, S., C. Plumhans, F.F. Behrendt, S. Stanzel, M. Suehling, G. Muhlenbruch, A.H. Mahnken, R.W. Gunther, and M. Das, *Semi-automated quantification of hepatic lesions in a phantom.* Invest Radiol, 2009. **44**(2): p. 82-8.
40. Buckler, A.J., L. Bresolin, N.R. Dunnick, and D.C. Sullivan, *A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging.* Radiology, 2011. **258**(3): p. 906-14.
41. Gavrielides, M.A., L.M. Kinnard, K.J. Myers, J. Peregoy, W.F. Pritchard, R. Zeng, J. Esparza, J. Karanian, and N. Petrick, *A resource for the development of methodologies for lung nodule size estimation: database of thoracic CT scans of an anthropomorphic phantom.* Optics Express, 2010. **18**(4): p. 15244-15255.
42. 3A-Working-Group. *Study 3A: Inter-method Study with Test-retest Clincal Data: Study Design Second Challenge* 2013; Available from: http://qibawiki.rsna.org/images/7/7b/3A_study_design%2C_second_challenge%2C_0.3.PDF.
43. *QI-Bench, free and open-source informatics tooling used to characterize the performance of quantitative medical imaging*. Available from: http://www.qi-bench.org/, accessed June 30, 2013.
44. CT-Volumetry-Technical-Committee. *QIBA Profile: CT Tumor Volume Change v2.2 Reviewed Draft (Publicly Reviewed Version)* 2012; Available from: http://rsna.org/uploadedFiles/RSNA/Content/Science_and_Education/QIBA/QIBA-CT%20Vol-TumorVolumeChangeProfile_v2.2_ReviewedDraft_08AUG2012.pdf.
45. Zhao, B., L.P. James, C.S. Moskowitz, P. Guo, et al., *Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer.* Radiology, 2009. **252**(1): p. 263-72.
46. Bland, J.M. and D.G. Altman, *Statistical methods for assessing agreement between two methods of clinical measurement.* Lancet, 1986. **1**(8476): p. 307-10.

47. Bland, J.M. and D.G. Altman, *Measuring agreement in method comparison studies.* Statistical Methods in Medical Research, 1999. **8**(2): p. 135-160.
48. Lin, L.I., *A concordance correlation coefficient to evaluate reproducibility.* Biometrics, 1989. **45**(1): p. 255-68.
49. Barnhart, H.X. and D.P. Barboriak, *Applications of the repeatability of quantitative imaging biomarkers: A review of statistical analysis of repeat data sets.* Translational Oncology, 2009. **2**(4): p. 231-235.
50. Warfield, S.K., K.H. Zou, and W.M. Wells, *Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation.* IEEE Trans Med Imaging, 2004. **23**(7): p. 903-21.
51. Rohlfing, T., D.B. Russakoff, and C.R. Maurer, Jr., *Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation.* IEEE Trans Med Imaging, 2004. **23**(8): p. 983-94.
52. Jaccard, P., *The distribution of the flora in the alpine zone.* New Phytologist, 1912. **11**: p. 37-50.
53. Sorensen, R., *A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons.* Nord Med, 1948. **40**(51): p. 2389.
54. Dice, L., *Measures of the Amount of Ecologic Association Between Species.* Ecology, 1945. **26**(3): p. 297-302.